王志华 Carey

15885844762 · wzh2657751462@gmail.com
个人主页：https://kirito14it.github.io/ · GitHub 主页：https://github.com/Kirito14IT

2023 级 | 仓库 star 总数约 2.5k | ACM 集训队员（银奖） | 国家励志奖学金 *5

教育经历

复旦大学 复旦计算与智能创新学院 电子信息-计算机（**直硕**） 2027.09–2030.07

课题组：**FVL**、**TEAI**（负责人：**姜育刚副校长**） 导师：**马兴军** 2026.03–至今

研究兴趣：LLM Coding Agent Agentic Training TrustWorthy AI | 英语 CET-6/CET-4：516/522

上海理工大学（本科生） 光电信息与计算机工程学院 **计算机科学与技术（院士创新班） 排名：3/85** 2023.09–2027.07

学术论文与知识产权

[1] **Zhihua Wang**, Nan Li, ..., Yunhao Feng, Songyang Wu. *ScamWeb: A Multimodal Benchmark for Grounded Understanding of Cyber-enabled Fraud Webpages*. Preprint; AAAI 2027（CCF-A）; Will be accepted(accept*4) **[First Author]**

[2] **Zhihua Wang**, Yunhao Feng, Yifeng Gao, Rui Pan, Baiheng Zhao, Zixing Chen, Xingjun Ma. *MemHarmBench: Evaluating Cross-Session Prompt Injection in Memory-Persistent AI Agents*. Preprint; WWW 2027（CCF-A） **[First Author]**

[3] **Zhihua Wang**, Yifeng Gao, ..., Xingjun Ma. *Smooth Inliers, Erratic Outliers: Frequency-Aware Mamba for Two-View Correspondence Learning*. Preprint; IEEE TIP（中科院一区 Top，CCF-A） **[First Author]**

[4] **Zhihua Wang**, Junfan Chen, ..., Guoxiang Tong. *SafeCodeRL: Security-Constrained Multi-Agent Reinforcement Learning for Trustworthy LLM-Generated IoT/CPS Software*. Sensors, 26(11), 3502 Accept（CCF-C） **[First Author]**

软件著作权：《DL 驱动的鲁棒型数字水印网络应用》《基于 CodeQL 与 LLM 的代码安全分析系统》**第一作者**

相关专利：信息网络安全公安部重点实验室（公安部第三研究所）相关专利申请正在评审。

研究经历

MemHarmBench：面向持久记忆投毒的跨会话 LLM Agent 安全基准 2026.07–2026.10

复旦大学计算与智能创新学院**姜育刚** & **马兴军**导师访问学生

上海创智学院推荐论文 | 奇绩创坛 Alpha Sight 推荐论文 | **WWW 2027** | **第一作者**

- 构建“暴露–触发”跨会话记忆投毒基准，覆盖 **288** 个案例环境（**24** 个攻击模板 × **12** 个业务场景）、**12** 类目标。
- 在恶意规则留存样本中，跨会话间接攻击成功率达 **59.38%–76.39%**，揭示正常执行下的决策操纵风险。

论文：https://arxiv.org/abs/2611.01700 **代码：**https://github.com/Kirito14IT/MemHarmBench **数据：**https://huggingface.co/datasets/kirco567/MemHarmBench

ScamWeb：面向网络诈骗网页检测、细分类与证据定位的多模态基准 2026.04–2026.08

复旦大学可信具身智能研究院**马兴军** & **姜育刚**老师实习学生 | AAAI 2027（accept*4） | **第一作者**

- 构建 **10,026** 页、**25** 类诈骗、**47** 种语言的多模态基准；提出免训练 ScamDet，支持检测、细分类与证据定位。
- 二分类 F1 达 **78.7%**；在 **3,324** 页诈骗子集的零样本评测中，召回率 **97.3%**、细分类准确率 **63.9%**。

论文 代码：https://github.com/Kirito14IT/ScamWeb **数据：**https://huggingface.co/datasets/kirco567/ScamWeb-Dataset **项目主页**

XSafeClaw：Agent 运行时安全治理的可信防护框架 Stars 165 获「机器之心」报道 2026.03–2026.06

复旦大学可信具身智能研究院 | 导师**马兴军** | 登上可信具身智能研究院首页“最新研究” | **核心贡献者**

- 面向 OpenClaw、Hermes 与 nanobot 提供会话风险评估、工具拦截和人工审批，支持多类 Agent 运行时安全防护。
- 主导 nanobot 接入，完成会话身份与工具事件统一、并发通信及三类运行时 Token/成本核算，完善防护与集成测试。

项目网站：https://xsafeclaw.ai **代码：**https://github.com/XSafeAI/XSafeClaw **报道：**https://mp.weixin.qq.com/s/1Dp5uTasxnm3ues_hNPlWQ

SFMambaNet：谱频增强选择性状态空间对应剪枝模型 2025.09–2026.02

福州大学福建省网络计算与智能信息处理重点实验室 | 导师：刘翼章 | IEEE TIP（中科院一区 Top，CCF-A） | **第一作者**

- 设计局部谱几何注意力与频域门控双向 Mamba，以近线性复杂度（$O(N \log N)$）抑制高频外点噪声。
- YFCC100M 未知场景（SIFT）mAP5° 达 **73.83**，较 MatchMamba 提升 **6.23** 个百分点；参数量 **2.01M**。

论文：https://arxiv.org/abs/2606.04493 **代码：**https://github.com/Kirito14IT/SFMambaNet **项目主页：**https://kirito14it.github.io/SFMambaNet/

实习经历

信息网络安全公安部重点实验室 (公安部第三研究所上海) AI 研究实习生 2026.04–至今

导师：吴松洋 **项目 1：**独立构建网页图文与 URL 联合分析及诈骗证据定位流程，完成**独立一作论文 1 篇**。

项目 2：设计**轻量级短文本风险主题发现算法**，融合字符 n-gram 与预训练多语言 MiniLM 编码器特征，用逻辑回归识别消息风险并排序会话；结合语义聚类和原文短语抽取，整理群聊主题与可追溯证据，支持本地 CPU 筛查。

国际组织（联合国志愿人员组织）–UNDP 数据研究实习生（线上） 2026.01–2026.04

导师：Johnny **项目：**远程办公，负责团队已采集数据的清洗、整理、分析与可视化，形成研究报告。

竞赛经历

2025 ACM 上海站大学生程序设计竞赛	**银 奖**	2025.07
2025 年第十六届蓝桥杯全国软件和信息技术专业人才大赛全国总决赛 C/C++	**国家级二等奖**	2025.06
2025 年睿抗机器人开发者大赛编程赛道全国总决赛	**国家级一等奖**	2025.08

技能与服务

编程与工具： C/C++、Python；PyTorch/TensorFlow、vllm；Transformers、Codex、Claude Code；**NeurIPS2026 审稿人**